Name: Achyut Shukla Batch: ML - 1 PRN: 20070122005

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act 1956)

Re - accredited by NAAC with 'A' Grade

Founder: Prof.Dr. S. B. Mujumdar, M.Sc., Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)

Lab Assignment — 5

Aim :

Apply Naive Bayes Classifier algorithm on a sample case study and data set. Evaluate Results.

PART — A

Theory :

1. Explain working of Naive Bayes Classifier with Bayes theorem in detail. Solve example to justify your answer.

   The Naive Bayes Classifier is a probabilistic machine learning algorithm used for classification tasks. It's based on the principles of Bayes' Theorem and assumes that the features are conditionally independent given the class label. Despite this "naive" assumption, Naive Bayes classifiers often perform surprisingly well in practice, especially for text classification tasks.

   Bayes' Theorem:

   Bayes' Theorem is a fundamental concept in probability theory that describes the probability of an event occurring, given the probability of another related event. Mathematically, Bayes' Theorem is expressed as follows:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

2. List types and explain.

   i) **Gaussian Naive Bayes:** This type of Naive Bayes assumes that the features follow a Gaussian (normal) distribution. It's suitable for continuous numerical features. In Gaussian Naive Bayes, the probability distributions for each class are modeled using mean and standard deviation parameters.
   ii) **Multinomial Naive Bayes:** Multinomial Naive Bayes is often used for text classification problems, where the features represent counts or frequencies of words or other discrete data. It's particularly effective for handling data with multiple categories and features that represent discrete counts.

iii) **Bernoulli Naive Bayes:** Bernoulli Naive Bayes is also used for text classification, specifically when the features are binary (i.e., presence or absence of a particular term). It's suitable for cases where you're dealing with binary data or want to model the absence or presence of features.

iv) **Complement Naive Bayes:** Complement Naive Bayes is a variation that was proposed to address the issue of imbalanced class distributions in text classification. It's designed to balance out the class distribution by using the complement of the class probabilities when making predictions.

v) **Categorical Naive Bayes:** Categorical Naive Bayes is used for categorical data where features have discrete categories. It can handle cases where the data isn't continuous or binary, but rather consists of distinct categories.

3. Strengths and Weaknesses of Naive Bayes Classifier.

Strengths of Naive Bayes Classifier:

1. **Simplicity and Speed:** Naive Bayes is a simple algorithm that is easy to implement and computationally efficient. It's particularly well-suited for large datasets and real-time applications.
2. **Scalability:** Naive Bayes can handle a large number of features without suffering from the "curse of dimensionality," making it useful for high-dimensional data.
3. **Decent Performance:** Despite its assumptions, Naive Bayes often performs surprisingly well in practice, especially for text classification and spam filtering tasks. It can achieve competitive accuracy with more complex algorithms.
4. **Handles Irrelevant Features:** Naive Bayes is robust to irrelevant features because it calculates probabilities independently for each feature given the class label.
5. **Effective for Text Data:** Naive Bayes works well for natural language processing tasks where features are often binary or categorical, such as text classification or sentiment analysis.
6. **Works with Small Training Data:** Naive Bayes can perform well even when you have a relatively small amount of training data, making it useful in cases where collecting a large dataset might be challenging.

Weaknesses of Naive Bayes Classifier:

1. **Strong Independence Assumption:** The "naive" assumption that features are conditionally independent given the class label might not hold in real-world scenarios. This can lead to suboptimal results if there are strong correlations among features.
2. **Limited Expressiveness:** Due to the assumption of feature independence, Naive Bayes might not capture complex relationships and interactions among features, which could lead to poor performance on certain datasets.
3. **Sensitive to Input Data:** Naive Bayes can be sensitive to small changes in input data, which might result in significant changes in the predicted probabilities and, consequently, the classification decision.
4. **Can't Handle Missing Data Well:** The presence of missing data can pose challenges for Naive Bayes, as it requires complete data to calculate probabilities accurately.
5. **Inadequate for Numerical Data:** While Gaussian Naive Bayes works for continuous numerical features, it assumes that the features follow a Gaussian distribution, which might not be true for all datasets.

6. **Class Imbalance:** Naive Bayes can struggle with imbalanced class distributions, especially when there's a significant difference in the number of samples between classes. This can lead to biased predictions.
7. **Limited Learning of Complex Patterns:** Naive Bayes is not well-suited for learning intricate relationships between features and class labels, as it doesn't capture more sophisticated patterns that might be present in the data.

PART — B

Experiment: Apply BN classifier algorithm on a sample case study and a dataset.

Steps:

- Data Pre-processing step
- Fitting Naive Bayes to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.

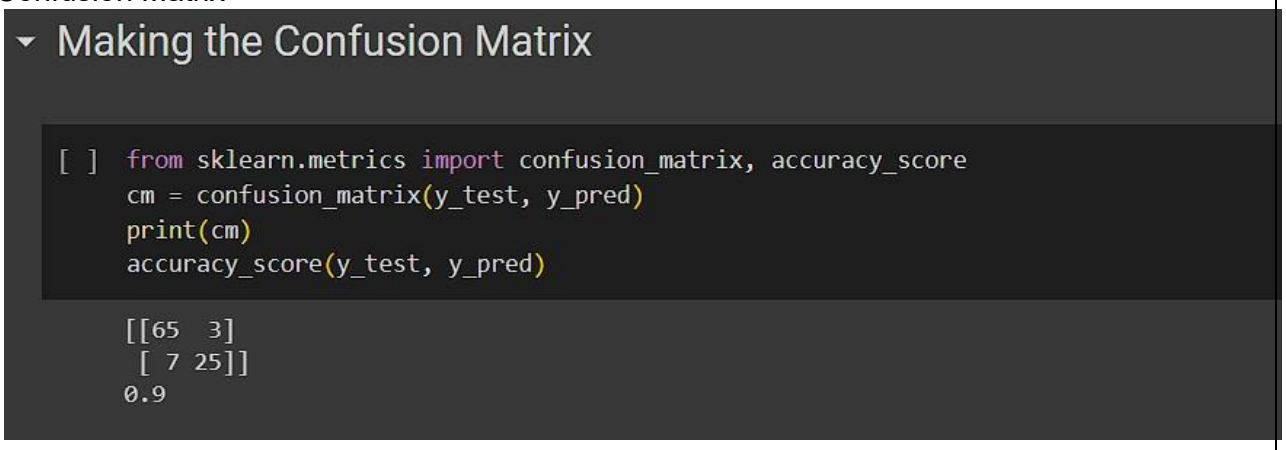
Implement Naive Bayes Classifier Algorithm and evaluate the results.

References :

1. https://github.com/gbroques/naive-bayes
2. https://github.com/topics/naive-bayes-classifier
3. https://towardsdatascience.com/naive-bayes-c1assifier-81d512f50a7c
4. https://www.kaggle.com/code/manan5598/decison-tree-and-naive-bayes-classifier

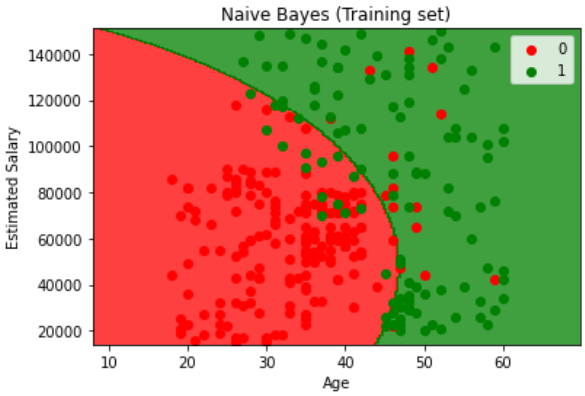
Output:

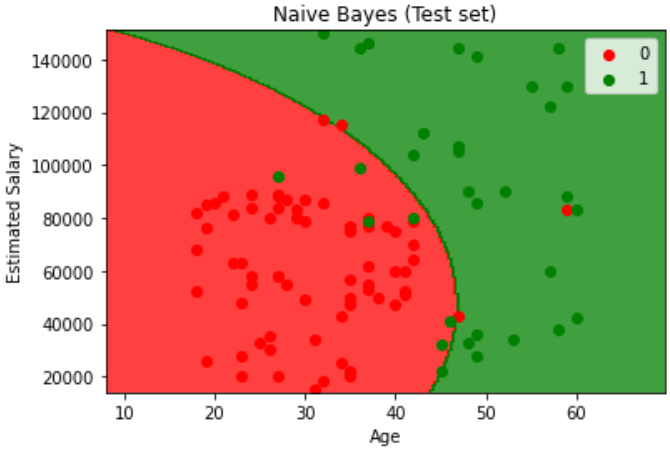
1. Confusion Matrix



Name: Achyut Shukla Batch: ML - 1 PRN: 20070122005

2. Data Visualization





Inference Discussion: Applied Naïve Bayes Classifier with an accuracy of 90%.